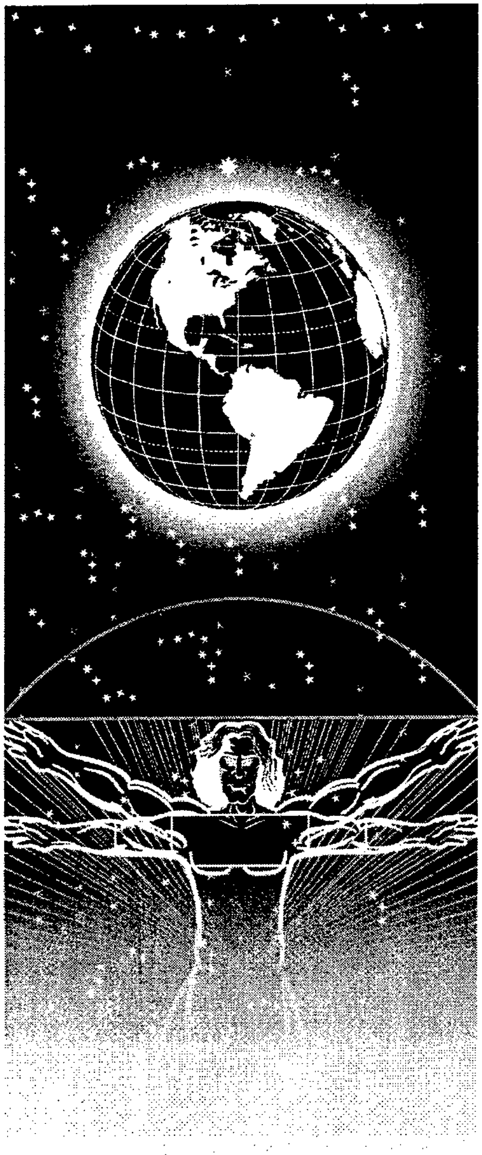
AFRL-HE-WP-SR-2000-0010



UNITED STATES AIR FORCE
RESEARCH LABORATORY

MODELING MENTAL WORKLOAD

Roger W. Schvaneveldt

NEW MEXICO STATE UNIVERSITY
COMPUTING RESEARCH LABORATORY
BOX 30001, MSC 3CRL
LAS CRUCES NM 88003

Gary B. Reid

HUMAN EFFECTIVENESS DIRECTORATE
CREW SYSTEM INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433-7022

Rebecca L. Gomez
Sean Rice

NEW MEXICO STATE UNIVERSITY
COMPUTING RESEARCH LABORATORY
BOX 30001, MSC 3CRL
LAS CRUCES NM 88003

SEPTEMBER 1997

FINAL REPORT FOR THE PERIOD 12 FEBRUARY 1996 TO 30 SEPTEMBER 1997

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

DISCLAIMER

This Special Report is published as received and has not been edited by the Air Force Research Laboratory, Human Effectiveness Directorate.

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-SR-2000-0010

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

MARIS M. VIKMANIS
Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 1997

3. REPORT TYPE AND DATES COVERED
Final Report, 12 Feb 1996 to 30 Sep 1997

4. TITLE AND SUBTITLE

Modeling Mental Workload

5. FUNDING NUMBERS

C F41624-96-1-0003
PE 62202F
PR 7184
TA 14
WU BA

6. AUTHOR(S)

Roger W. Schvaneveldt* Sean Rice*
Gary B. Reid**
Rebecca L. Gomez*

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

*New Mexico State University
Computing Research Laboratory
Box 30001, MSC 3CRL
Las Cruces NM 88003

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory**
Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFRL-HE-WP-SR-2000-0010

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The primary objective of the research project was to investigate models for monitoring and predicting subjective workload in the control of complex systems. Such models would enable systems to use workload levels to distribute tasks optimally in addition to identifying levels of workload, which could lead to a serious breakdown in performance. In the aircraft-pilot system, for example, such capabilities could provide warnings to the pilot of high workload levels and could also assess ways of reducing the pilot's workload by offering to assume control of some ongoing tasks. In this initial project, we tried to determine how well a model can assess workload using information about task requirements and task performance.

14. SUBJECT TERMS

Modeling, Complex Systems, Workload, Pilot, Aircraft

15. NUMBER OF PAGES
41

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Abstract

The primary objective of this research project was to investigate models for monitoring and predicting subjective workload in the control of complex systems. Such models would enable systems to use workload levels to distribute tasks optimally in addition to identifying levels of workload which could lead to a serious breakdown in performance. In the aircraft-pilot system, for example, such capabilities could provide warnings to the pilot of high workload levels and could also assess ways of reducing the pilot's workload by offering to assume control of some ongoing tasks. In this initial project, we tried to determine how well a model can assess workload using information about task requirements and task performance.

Participants rated subjective workload levels after each block of trials. The blocks consisted of various combinations of three tasks with varying levels of difficulty. The workload ratings and the performance data were used to create a database for developing models. The tasks were: (a) a continuous tracking task with a random forcing function and three different updating speeds; (b) a discrete tracking task in which response keys were pressed to indicate the position of a target in one of four different locations; and (c) a tone-counting task which required counting the number of higher pitched tones in a series of tones of 800 or 1200 Hz. Neural net models applied to group data consisting of eight individuals were able to achieve 85-95% accuracy in predicting a "redline" workload level in training data. On completely new data, accuracy was in the 70-75% range. The redline value was adopted from earlier work (Reid & Colle, 1988) showing that at that value of workload, performance measures begin to show effects of workload. Linear models (no hidden units) performed about as well as nonlinear ones in prediction using new data. Thus, for the cases we studied, linear regression models would do as well as nonlinear neural network models.

Prediction in the 70-75% range is of interest theoretically, but for practical utility, values in the range of 90-95% are desirable. When we developed models from the data of individual participants, such levels were reached for three of eight participants, but the other six were in the 74-85% range. The average accuracy from individual participant models was better than that obtained with group data suggesting that individual models are a more promising direction to pursue. We conclude with a discussion of possible directions to pursue in attaining greater levels of accuracy.

TABLE OF CONTENTS

Introduction

Mental workload is a multi-faceted phenomenon, and the literature reflects these many facets. Mental workload can be related to physiological states of stress and effort, to subjective experiences of stress, mental effort, and time pressure, and to objective measures of performance levels and to breakdown in performance. These various aspects of workload have led to distinct means for assessing workload including physiological criteria (e.g., heart rate, evoked potentials), performance criteria (e.g., quantity and quality of performance), and subjective criteria (e.g., ratings of level of effort).

According to performance criteria, a given task will not necessarily lead to a particular level of performance or workload because factors such as S-R compatibility, practice, fatigue, talent or skill, etc. will affect task workload. For example, a task which may seem overwhelming when first attempted may end up requiring only a small amount of mental capacity after sufficient practice. People learning to fly commonly experience a dramatic reduction in the workload imposed in landing after extended practice. Several aspects of the environment and the aircraft must be monitored and controlled in executing the approach and landing. Particularly at low levels of practice and familiarity, these many aspects of monitoring and control can easily exceed a person's information processing capacity[1].

Rogers and Monsell (1995) have shown persistent costs associated with switching between tasks even when the switches are predictable and regular. Schvaneveldt (1969) showed that performance on relatively simple tasks can be degraded when they are coupled with complex, independent tasks. Moray, Dessouky, Kijowski, & Adapathya (1991) showed clear limits to performance in the context of scheduling multiple tasks. Thus, there is reason to believe that the

[1] Practice can reduce workload, but some studies (Schneider & Detweiler, 1988; Schvaneveldt & Gomez, 1998) suggest that practice on single tasks leads to rather poor transfer to dual task situations

requirement to perform multiple tasks is a major contributor to performance levels and, as a result, to workload (Wickens & Yeh, 1982).

The literature on mental workload was extensively reviewed in two chapters in the 1986 *Handbook of Perception and Human Performance*. On the theoretical side of the problem, understanding workload relates primarily to research in attention, processing capacity, dual-task performance, and allocation of mental resources (Gopher & Donchin, 1986). Assessing workload has involved measurement of performance, subjective impressions of workload, and physiological indicators of work and stress (O'Donnell & Eggemeier; 1986). Because subjective measures of workload have proven useful in a variety of circumstances, we decided to concentrate on these measures in the present investigation.

In regard to subjective measures, it is likely that people do not have conscious access to all aspects of mental workload which may cause particular difficulty with subjective measures. Despite this limitation, several studies attest to the value of such measures (Bortolussi, Kantowitz, & Hart, 1986; Corwin, 1992; Haskell & Reid, 1987; Reid & Colle, 1988; Tsang & Vidulich, 1994; Vidulich, Ward, & Schueren, 1991; Wierwille & Eggemeier, 1993). It is also important to consider that subjective workload represents the degree to which an individual experiences workload demands, and this experience itself has potential consequences for performance and stress levels. Thus for both theoretical and practical reasons, it is of value to characterize how much mental effort is experienced in performing various tasks and to predict when performance will deteriorate seriously due to overload.

Subjective Measures of Mental Workload

Subjective measures of mental workload are obtained from direct estimates of task difficulty. Various techniques can be used to measure subjective workload, but the basis of any subjective workload technique is having participants report about the "difficulty" of the task. The main difference between subjective workload and other workload measures (such as dual-task or physiological measures of workload) is that the former rely on the participants' conscious, perceived experience with regard to the interaction between the operator and the system.

Techniques for assessing subjective workload fall either in the category of ratings scales procedures or psychometric techniques involving such procedures as magnitude estimation (e.g. Borg, 1978), paired comparisons (e.g. Wolfe, 1978), or conjoint measurement and scaling (e.g. Reid, Shingledecker, & Eggemeier, 1981). Ratings procedures such as those derived from the Cooper-Harper Aircraft-Handling Characteristics Scale (Cooper & Harper, 1969) require participants to rate the difficulty of tasks with the use of a decision tree. Ratings scales appear to be sensitive to different levels and varieties of load (including perceptual, central processing, and communications load). An advantage of psychometric over ratings techniques, is that the former are capable of providing interval information regarding task difficulty. Such information can be useful for measuring the magnitude of workload differences between tasks.

In magnitude estimation, participants provide direct estimates of the difficulty of one task relative to another. For example, participants are exposed to a task and then are told to choose a numerical value reflecting the difficulty associated with the task. The perception of difficulty associated with the first task is called the modulus. Participants are then asked to provide numerical estimates of tasks of varying difficulty relative to the modulus. Magnitude estimation has proven to be a sensitive measure of differences in load but its major drawback is that realworld tasks often do not occur in close proximity, thus making it difficult for participants to retain an accurate representation of the modulus over time. In paired-comparisons, participants are presented with all possible pairs of stimuli (e.g., difficulty levels of a task) and are asked to judge which of the two stimuli are more difficult. After comparisons are obtained from a number of participants the relative difficulty of stimuli can be represented in an *n* x *n* matrix showing the proportion of times each stimulus was judged to be more difficult than every other stimulus. Although this technique has also produced successful results, the number of comparisons required is a limiting factor. For example, with 6 stimuli, 15 judgments are required.

An approach which has proven useful is the technique of conjoint measurement and scaling. Most of the techniques used for measuring subjective workload treat perceived workload as a unitary dimension. In some cases participants are asked to consider multiple factors in making a rating, but participants still assign one number based on these factors. However, subjective workload entails a number of dimensions, such as time load, mental effort load, and psychological stress load (Reid et al., 1981). Conjoint measurement and scaling approaches are multidimensional techniques with the advantage of reflecting a number of factors in one measure of subjective workload. This approach involves obtaining separate ordinal ratings for each of several dimension of subjective workload and then combining the separate ratings into a scale with interval properties. These techniques require two phases: a scale development phase and an event scoring phase. During the scale development phase, levels of dimensions are described to participants. Then participants are given all combinations of descriptions of each of the levels for the dimensions and are asked to rank order the combinations according to workload. If there are three dimensions and three levels of difficulty, then participants would rate 27 combinations. The rankings are then submitted to a series of axiom tests which are part of the conjoint measurement procedure. These axioms are used to test logical consistencies in the data and identify the combination rule (e.g., additive, distributive, dual distributive) that fits the data of a given participant. The rule is then used to assign numerical values to each level of the separate dimensions and then combine the values into one integrated scale. During the event-scoring phase, participants participate in a task and then rate the task difficulty on each of the dimensions. The ratings are then used to find the corresponding value on the participant's interval scale. Conjoint measures appear to be sensitive to levels of task difficulty. Additional advantages of this approach are that measures are easy to obtain and can be scaled individually to participants.

Participants are very consistent in their ratings with subjective measures of workload. Reliability coefficients for subjective workload over multiple ratings instances have been as high or higher than .90 (Gopher & Browne, 1984). However, the relationship between subjective and objective measures of workload is variable. In some instances researchers report an association between subjective and objective measures of workload, in other instances dissociations are reported. One explanation for these inconsistencies has to do with the relationship between processes which are and are not available to consciousness. On this view, subjective workload measures will be more sensitive to processes which require awareness (or attention) and less sensitive to processes which do not require attention. According to Gopher and Donchin (1986) the retrospective nature of subjective workload may also be a contributor to the dissociations between subjective and objective measures. Regardless of the limits of subjective measures, the subjective experience of performing a task cannot be ignored. Often, subjective experiences of overload take precedence when an operator is performing a task, even when objective measures are not indicating an overload (Moray, Johanssen, Pew, Rasmussen, Sangers, & Wickens, 1979).

The Present Study

We used empirical, analytical, and computer modeling methods to investigate mental workload in the performance of system control tasks. Our specific objectives were: (1) to collect performance measures and subjective judgments of workload using pursuit tracking, tone counting, and sequential reaction-time tasks under conditions which varied the number and complexity of tasks to be performed; and (2) to develop models for predicting subjective workload judgments from performance measures and task conditions. Such models can be used to monitor levels of workload for the purpose of predicting when performance will seriously deteriorate due to overload and to analyze the impact of allocating various tasks to an automated system. The following sections discuss the tasks we used in our study and relevant literature is reviewed.

Tracking Tasks

We used a pursuit tracking task as one component of our workload study. Tracking is representative of a central task involved in flying aircraft. Such tasks also have the advantage of providing continuous measures of performance, and the difficulty of the task can be easily manipulated by variations in forcing functions (Wickens, 1986). We conducted some preliminary studies to select appropriate parameters of the tracking task in order to tune the difficulty of our tasks to the participant populations we studied. We found that reliable changes in workload could be achieved by varying the rate at which a random forcing function updated the position of the target. In the tracking tasks, we recorded tracking error and stick movement variance measures over blocks of trials. We assessed control activity by variability of stick movement.

Continuing our analogy with flight control. Flying also involves numerous other tasks (e.g., navigation, communication, engine control). It has been argued that a major contributor to workload is the number of tasks that must be performed simultaneously (Moray, Dessouky, Kijowski, & Adapathya, 1991; Rogers and Monsell, 1995; Schvaneveldt, 1969; Wickens & Yeh, 1982). Thus we selected additional tasks with particular qualities as described in the next section.

Sequential Reaction-Time Tasks

The sequential reaction time (SRT) task provided an interesting dual task with tracking because the SRT task can be manipulated in structural complexity (by varying constraints in the sequence of events). It has been known for some time that people are sensitive to the sequential order of events in reaction-time tasks (Schvaneveldt & Chase, 1969). With first order constraints, an event is predictable knowing only the most recent event. Second order constraints require knowledge of the preceding two events to predict the next event. Hyman (1953) showed that reaction time varies with the probability of events when either overall or sequential probability is manipulated.

The approach most often taken to dual-task manipulations of workload is to assess the effect of a secondary task such as Sternberg memory-scanning tasks (Wickens, Hyman, Dellinger, Taylor, & Meador, 1986) or scheduling tasks (Moray, et al., 1991) on a continuous primary task (such as tracking). One difficulty with this approach is that there is no way of ensuring that participants will treat the primary task as primary and the secondary task as secondary. Therefore, a more realistic approach, and the one we chose to follow, was to interpret the contributions multiple tasks made to the complexity and hence to subjective workload involved in controlling the system.

The SRT task we used required participants to respond to the position of "blots" on the screen by pressing one of four response keys indicating which of four positions contained the blot on each of a series of trials. Figure 1 shows the display we used with the blot on the third bar from the left. The correct response would be to press the third key from the left.

Insert Figure 1 about here

During the SRT task, participants are exposed to event sequences with a repeatable pattern. In most variants of the SRT task, a target occurs in one of three to six locations as dictated by a pattern sequence. The complexity of the stimulus pattern can be varied by manipulating the order of sequential constraint. That is, the stimulus pattern is designed so that predicting the next event depends on one, two, or more preceding events. Learning of the sequence structure is measured by the disparity in reaction-times for responding in the structured sequence in comparison to some change in the sequence. For example, practice with a structured sequence results in a dramatic decrease in reaction time as compared to the reaction time for responding to randomly generated sequence locations (Cohen, Ivry, & Keele, 1990; Nissen & Bullemer, 1987).

A number of researchers have investigated the relationship between workload and performance on the SRT task (Cohen et al., 1990; Curran & Keele, 1993; Frensch, Buchner & Lin, 1994; Nissen & Bullemer, 1987; Perruchet & Amorim, 1992; Reed & Johnson, 1994; Schvaneveldt & Gomez, 1996; 1998). Nissen and Bullemer (1987) exposed participants repeatedly to a 10-trial pattern in a SRT task. Asterisks appeared in one of four horizontal locations on the screen and participants pressed a corresponding response button. Nissen and Bullemer measured performance under normal, dual-task, and random conditions. Participants in the normal sequence condition demonstrated sensitivity to the structure of the sequence as evidenced by improved reaction times relative to participants receiving random sequences. In the dual-task condition high and low pitch tones accompanied the sequence learning task. Participants in this condition counted the number of low tones. Sensitivity to the structure of the sequence was severely impeded under dual-task conditions. Specifically, performance was no better than was responding to the random sequence, suggesting that sensitivity to structure in the environment is dependent on attentional processing.

In another important study of the relationship between sequential reaction-time and workload, Curran and Keele (1993) investigated the hypothesis that humans have two independent mechanisms for exhibiting sensitivity to sequential structure. One mechanism requires attention to the relationship between successive stimulus events whereas the other mechanism requires no such attention. The relationship between these two mechanisms was explored by assessing transfer of sensitivity to structured sequences under varying conditions of workload. When participants participated in an SRT task under single task conditions both attentional and nonattentional mechanisms operated in parallel. However under conditions of increased workload, the attentional mechanism was disabled, while leaving the nonattentional mechanism intact. Furthermore, the nonattentional mechanism shows sensitivity to simple first-order conditional structure and hybrid structure (combinations of first-order and second-order relations), but does not appear to be sensitive to more complex structure (such as that made up entirely of second-order conditionals). Such a finding is important because it suggests that certain tasks may be impervious to conditions of excessive workload, namely those which are processed by the nonattentional mechanism. It would be of potential use to identify tasks which are processed by the nonattentional mechanism because, presumably, these tasks would not deteriorate during system overload.

In summary, the tasks we used in our research reported here included the presence or absence of each of three tasks, pursuit tracking, sequential reaction time (SRT), and tone counting. The difficulty of tracking was varied by manipulating the speed of the cursor to be tracked, the SRT task difficulty was varied by using structured sequences vs. random sequences of blot positions. The tone task always accompanied the SRT task when the tone task was performed, but its presence was manipulated. In all there were 22 different conditions which were administered on each of 10 days to each of eight participants. The conditions varied in the combinations of tasks and the difficulty of the tasks.

In our studies, following each block of trials, participants gave three ratings of workload for that block of trials. The ratings used the SWAT scaling methodology (Reid, Potter, & Bressler, 1989). This method required participants to give 3-point ratings of each of three scales, time pressure, mental effort extended, and stress levels. These ratings in turn were scaled to produce workload values on a 0-100 scale. We converted these to a 0-1 scale to conform to our conventions for neural network model development. Next, we discuss our approach to modeling and predicting mental workload, namely neural network models.

Neural Network Models of Workload

We trained neural network (NN) models to predict subjective workload measures using both condition and performance measures as input variables. Essentially the models attempted to learn how condition and performance factors relate to differences in subjective workload measures.

There are several possible approaches to developing a neural net model of mental workload. Because of the extensive development and frequent application of the multi-layer feed-forward network architecture (or multi-layered perceptron, MLP), it is a reasonable place to start. With that architecture, it is possible to compare the performance of a linear system (a perceptron which has no hidden layers and can only represent linear solutions; see Rosenblatt, 1962) with the nonlinear systems that can be realized by including one or more layers of hidden units. If there is no performance gain with the inclusion of hidden units, the problem has a linear solution (or the best solution to the problem is linear). In such cases, the nonlinear solution with the hidden layer(s), tends to be unstable, and it generalizes to new cases poorly. In our modeling work, we routinely compared more complex models to the linear ones. To give the linear models a reasonable chance of performing well, it is important to code inputs such that the expected output has a monotonic relation with the input code. For example, in coding different conditions, the difficulty of the conditions should be represented by the order of values in the variable coding the conditions so, for example, coding the tracking task conditions might use 0, 0.33, 0.67, and 1.0 to code no tracking, and slow, medium, and fast cursor movement, respectively.

In contrast to rule-based models, NNs compute by using interconnected networks of simple processing units. These simple units, called nodes, receive information from external sources (i.e., from input to the network or from other nodes), sum this information, and then propagate an activation level to all connected nodes. The advocates of NNs frequently mention the ability of such systems to learn the appropriate mapping of inputs to outputs from examples and to successfully generalize that learning to new examples. Perhaps the crux of neural network modeling is the application of appropriate learning algorithms to appropriate processing network topologies such that a set of connection weights is found that lead to desired performance. One of the most basic learning algorithms found in NN models is the Hebbian contiguity, or associative rule (Hebb, 1949). This simple learning rule states that if two simple processors are simultaneously active and are connected, then the relationship between them should be strengthened. This type of learning rule is associated with networks that rely on external teachers for feedback. In such networks, learning occurs in an iterative feedback loop composed of four parts (Lippmann, 1987). First, a pattern is presented and activation is propagated through the layers of the network. Second, the output activation is compared against the correct output (i.e., the true output information associated with a given input pattern), and an error term is computed. Third, interconnections (i.e., weights) are modified using some scheme that reduces the error measure computed in part two. Finally, go back to step one and repeat this process. This iterative learning scheme continues until all training patterns produce the correct output.

NNs are able to generalize from previously learned responses to incomplete or novel instances of stimuli. To the extent that a new stimulus is similar to a stimulus pattern that has already been trained into the network, a similar pattern of processing will occur across the network resulting in a similar response (Arbib, 1986).

The multi-layer perceptron (MLP) model has one or more layers of processing nodes between input information and the output layer. A layer is a set of processing nodes connected to successive layer nodes via a matrix of weights. That is, a weight matrix represents the connectivity in a layer where the row dimension of the matrix corresponds inputs for a given layer and the column dimension represents outputs for that layer. The computational power of the multi-layer perceptron stems from the application of non-linear activation functions, as well as the associated family of non-linear learning algorithms such as the back propagation gradient descent (Rumelhart, Hinton, & Williams, 1986; Rumelhart, & Zipser, 1986).

The operation of a feed-forward network operates by passing activation from the inputs for each layer to the outputs in the layer via the weights on the connections between the inputs and the outputs. The net input to a given node is passed through a non-linear quashing function which keeps the activation of a unit between zero and one. This passing of activation through the layers is repeated until the final outputs are computed in this way. Learning in such networks is a matter of finding a set of weights which will compute a desired input-output mapping. If the mapping is linear, a single layer is sufficient. With nonlinear mappings, "hidden nodes" are required resulting in at least two layers in the system.

Learning with hidden layers requires using back propagation to modify the weights in the network. Back propagation is termed a gradient-descent method in that a measure of error for every weight in the network is being reduced by the learning algorithm. One way of thinking of this is that there is some n-dimensional space where the weights that define a network at a given time reside. The weights can be thought of as representing a surface in this space. To the extent that input patterns are incorrectly classified (i.e., produce inappropriate output activation) then there is error associated with the weights that define the surface in the n-dimensional space. The back propagation algorithm attempts to minimize this error by modifying the weight space (Plaut, Nowlan, & Hinton, 1986). When a network has been trained to classify a set of stimulus patterns the weight space that provides the solution is said to be at a minimum in the sense that the error associated with the weight surface is minimized. That is, the error found in all the weights has been minimized such that all training set input patterns are transformed by the weight layers resulting in correct classifications. Just as with any gradient-descent method, the back propagation procedure is subject to becoming trapped in a "local minimum" and, consequently, failing to find the best solution for a problem. With some problems, the local minima may be numerous causing great difficulty with the learning of the input-output mapping.

This completes our background discussion of research and methods related to our work. The remainder of the paper discusses the specifics of the work we performed.

Research objectives.2 Program objectives.

The long-range objective of this research was to develop a model for monitoring workload in complex systems. Such a model would enable systems to use workload levels to distribute tasks optimally. In the aircraft-pilot system, for example, such capabilities could provide warnings to the pilot of high workload levels and could also assess ways of reducing the pilot's workload by offering to assume control of some ongoing tasks. Our goal was to determine how well a model can assess workload using data from performance and task situations. Our specific objectives were: (1) to collect performance measures and subjective judgments of workload from pursuit tracking and sequential reaction-time tasks under conditions which varied the number and complexity of tasks to be performed; and (2) to develop models for predicting subjective workload judgments from performance measures and task conditions.

Methods

Participants. Eight participants participated in one session per day over a period of 10 days.

Apparatus. The data were collected using a Gateway 2000 486 50mhz computer with interfaces for a joy stick and push buttons.

Description of Tasks, Procedure and Materials. Participants participated in four tasks: subjective workload assessment, a pursuit tracking task, a tone-counting task, and a sequential reaction-time task. The latter three tasks were performed under multiple task conditions.

Subjective Workload Assessment Task

SWAT, the subjective workload assessment technique (Reid, Potter, & Bressler, 1989) was used to measure subjective workload during the processing of multiple tasks. This technique measures three components of perceived mental workload: time-load, mental effort load, and psychological stress load. Each of the three components is further described in terms of three levels of load. The SWAT involves a two stage procedure. The purpose of the first stage is to develop individual scales reflecting each participant's internal model of workload. The purpose of the second stage is to score events in terms of subjective mental workload. During the scale assessment phase participants ordered a deck of 27 cards, where each card represented a combination of one of the three levels of the three factors contributing to subjective workload. Participants ordered the cards from the lowest imaginable combination of workload to the highest imaginable combination. Conjoint analysis was used to convert the card sorting data to a scale where zero represented the combination of factors with the lowest perceived mental workload and 100 represented the combination with the highest perceived mental workload. This scale was used to assign individualized, subjective workload ratings to tasks in the event-scoring phase.

During the event scoring phase, participants were periodically asked to rate the workload associated with a series of trials. Participants were instructed to use the same three level descriptors as they used in the first phase of the SWAT procedure to describe each of the three factors. The ratings combinations obtained in the event scoring phase were then mapped back onto the individualized scale values in order to assign a particular SWAT value to the event.

The first phase of the SWAT took approximately one-hour. The second phase occurred between blocks of the performance tasks and took no more than a few seconds for the three ratings.

Pursuit Tracking Task

In the pursuit tracking task, the participant attempted to keep a moving cursor (i.e., a cross) on a computer screen within the boundaries of a pursuit target (i.e., an open circle) by controlling the cursor with a one-axis joystick. The target moved to the right and left according to a random forcing function. The speed of the movement of the target was varied over three levels: slow, medium, and fast corresponding to update intervals of 400, 200, and 100 msec respectively. Preliminary studies showed that these intervals led to the desired variation in tracking difficulty. In the tracking tasks, we recorded sufficiently detailed data to allow RMS error (and other measures of performance) to be computed over varying time segments.

Sequential Reaction-time (SRT) Task

During the SRT task, a "blot" appeared in one of four horizontal positions in the center of the screen. Four vertical line segments demarcated the four spatial positions. The blot appearing in one of these locations essentially increased the size of the line at on of the locations. Participants made responses by pressing one of four response keys corresponding to the spatial location of the blot. The blot remained on the screen until participants made the correct response.

Two second-order conditional sequences were taken from Reed and Johnson (1994). Sequence A, a 12-item sequence was 1-2-1-3-4-2-3-1-4-3-2-4. Sequence B was 1-2-3-4-1-3-2-1-4-2-4-3. The sequences were equated with respect to frequency of location (each location occurred three times), number of reversals (e.g. 1-2-1, one for each sequence), first-order transitions (each location was preceded once by the other three locations), and repetitions (no repetitions in either sequence). The only difference between sequences was in second-order conditional structure. Structure of the sequence was manipulated so that in some blocks the sequence was random and in other blocks the sequence was structured. In random blocks both sequences occurred with a .5 probability. In structured sequences Sequence A occurred with a probability of 0.90 and Sequence B occurred with a probability of 0.10. The probabilistic sequences were implemented by using the last two events to select the next event. Thus with probability 0.90, the next event would be the event in the probable sequence following the two just preceding events, and with probability 0.10, the next event would come from the improbable sequence. The discrepancy in RT to respond to probable vs. improbable sequence location reflected sensitivity to the structure of the sequence. Each block of trials was started by randomly selecting two of the four events for the first two trials in the block. There were 50 trials in each block.

Tone-Counting Task

In the tone-counting task, which only occurred in the context of the SRT task (i.e., participants never participated in tone counting when tracking was the only task), a high- or low-pitched tone occurred 17 ms after the response in the SRT task. The pitch of the tone was 800 or 1200 Hz. For each block of trials, the probability of a high-pitched tone was randomly set in the range of 0.40 to 0.60. The participant's task was to count the number of high-pitched tones in the block of trials and report the number at the end of the block.

Combining Tasks

The open circle participants followed in the tracking task moved in a horizontal line approximately 1/4 inch above the horizontal bars displayed in the center of the screen for the SRT task. In this way it was possible to keep information for both tasks in the participant's field of vision. Both the tracking task and the SRT task were performed alone in some blocks of trials. The tracking task was performed with the SRT task and with both the SRT and tone-counting tasks. In another combination, the SRT task was performed with just the tone-counting task. In all these cases, the difficulty of the tracking task and of the SRT task were also manipulated. The tracking task was slow, medium, or fast. The SRT task involved structured or random sequences of stimuli.

Subjective Workload Ratings

After a participant completed a block of 50 trials in the SRT task and/or approximately 35 seconds of tracking, the program prompted participants for three ratings of subjective workload each on a three point scale. Participants entered three numbers between 1 and 3. The first rating was for time load, the second for mental effort load, and the third rating was for psychological stress load.

Feedback

Feedback was displayed at the end of each block. Feedback for the tracking task consisted of the distance between the target and the cursor sampled every 100 msec and averaged over approximately 35 seconds; feedback for the SRT task consisted of mean reaction time and accuracy; and feedback for the tone-counting task consisted of displaying the number of high-pitched tones reported by the participant and the actual number of high-pitched tones. Participants were allowed to rest as long as desired between blocks.

Procedure

In the first session, participants received instructions regarding the number of sessions and the tasks involved in each session. Participants were told that the purpose of the experiment was to learn more about the effect of practice on motor performance. Participants also participated in the 22 blocks of trials representing the various combinations of conditions. Then participants participated in constructing the SWAT rating scale. Participants used initial exposure to the various conditions as a reference in establishing their SWAT scales. A schedule for the remaining 9 sessions was then arranged. Participants were encouraged to participate in one session per day on a daily basis. Each session consisted of 22 blocks representing the various combinations of conditions. The computer program prompted participants for subjective workload ratings following each block of trials.

Participants were paid $100 each for their assistance in the project. In addition, incentives for good performance were offered in the form of a $100 bonus for the individual with the best overall performance taking all of the performance variables into account using standardized scores. A $50 bonus was paid to the individual with the second best overall performance.

Neural Networks

The system we used for training and testing NN models was implemented by New Mexico State University (NMSU) and University of New Mexico (UNM) under earlier contracts with AL/HRA (Benson, Schvaneveldt, & Waag, 1994; Schvaneveldt, Benson, Goldsmith, & Waag, 1992). The model is a multi-layer perceptron (MLP) that may have one or more layers of processing nodes between input information and the output layer. The linear (perceptron) case is also handled by the software by simply not including any hidden layers. The model uses the back-propagation rule to learn the training patterns, and the trained model can be tested on new patterns not used in training. The NN models were trained using the performance data, the condition information, and the subjective judgments of workload previously described.

Results and Discussion

For a general source of information about all of the variables in the study, we present Table 1 which shows pairwise correlations of all of the variables. This table is primarily included for completeness because we discuss many of the relations among the variables in the following sections. Table 2 gives description of the variables shown in Table 1. These variables were the basis for the analyses we report and also the basis of the neural net analyses discussed later.

Insert Tables 1 & 2 about here

In reporting the results of our investigations, we first discuss the performance of the tracking, SRT, and tone counting tasks. Next we examine the effects of the tasks on workload. Finally, we discuss the neural network models designed to predict workload from information about conditions and/or task performance.

Task Performance

The manipulations of task difficulty both by varying the difficulty of a task itself and by adding additional tasks generally systematically affected task performance. The exception was tone counting performance which was not affected by the difficulty of the tracking task or the difficulty of the SRT task. The analyses presented here cover all of the performance data over days 3 through 10. Days 1 and 2 were assumed to reflect practice and start-up variability so they were not included. Tables 3 and 4 show mean performance measures for tracking error and SRT reaction time, respectively. In these tables, performance is shown on a relative (0-1) scale as the data were used in modeling. An “x” in the table indicates that a particular condition was not included or that the task in question was not performed in that condition. In Table 3, the last column shows that tracking error was clearly affected by the tracking condition (F(2,1149)=353.92, p<.001). It was also affected by adding the SRT task (F(1,1150)=126.02, p<.001). Adding tone counting to the SRT task also significantly increased track error (F(1,766)=3.94, p=.047). Thus, performance on tracking was affected by all of the task manipulations.

Insert Table 3 about here

In Table 4, the difference in reaction time between structured and random sequences is significant (F(1,1022)=4.21, p=.040). The tracking task also affects reaction time in the SRT task both when the analysis includes the absence of the tracking task (F(3,1020)=210.21, p<.001) and when only variation in difficulty of the tracking task is analyzed.(F(2,765)=14.20, p<.001). Adding tone counting also significantly increase SRT reaction time (F(1,1022)=10.60, p<.001). Thus, the SRT task also showed significant influences from variations in the presence and difficulty of the other tasks.

Insert Table 4 about here

Effects of Tasks on Workload

Now we turn our attention to the effects of the task manipulations on measures of subjective workload. As we had expected, the manipulation of the number of tasks and the difficulty of task components generally had significant effects on workload ratings. Table 5 shows the average SWAT scale values in the various conditions of the experiment across all of the data from days 3 through 10. The variation in workload with tracking conditions is statistically significant whether all conditions are compared (F(3,1404)=120.70, p < .001) or just comparing the variations in tracking difficulty without the No Tracking condition (F(2,1149)=99.47, p<.001). The effect of the SRT task condition is significant when No SRT task is included

(F(2,1405)=99.67, p<.001), but although the difference between structured and random SRT sequences is in the expected direction, it is not significant. The difference in workload with and without tones is significant (F(1,1406)=270.11, p<.001). Thus, the manipulation of the number of tasks and the difficulty of the tasks do generally affect workload levels. An examination of the means in Table 5 shows very orderly increases in workload as tasks are added and/or the difficulty of tasks is increased.

Insert Table 5 about here

Table 6 shows the analysis of workload using "redline" measures. The redline value was adopted from earlier work (Reid & Colle, 1988) showing that at values of workload around 40 on the SWAT scale (0.4 on our rescaled values), performance measures begin to show effects of workload. Thus it is useful to determine the extent to which various conditions lead to workload levels in excess of this redline value. In our neural network modeling, we came to focus on predicting whether workload exceeded redline or not because this discrete decision leads to clear interpretations of the outcome of modeling: how accurately can we predict workload levels over redline? We coded each trial block in terms of whether workload equaled or exceeded redline (value = 1) or not (value = 0). In our data set, the identical redline coding results from using either the Group SWAT Scale or the Individual SWAT Scales. When the 0 and 1 values are averaged over participants and days, the values show the proportion of trial blocks on which workload exceeded redline. These means are shown in Table 6. The statistical significance of differences in these data matches that for the data in Table 5. Variations in the tracking task leads to significant differences whether the comparison includes the no tracking condition (F(3,1404)=66.81, p < .001) or not (F(2,1149)=44.37, p<.001). Adding the SRT task increases workload (F(2,1405)=55.16, p<.001), but variation in the structure of the SRT sequence does not significantly affect workload. Adding tones significantly increases workload (F(1,1406)=158.19, p<.001). Inspection of Table 6 shows, as with Table 5, that workload increases systematically as tasks are added and/or the difficulty of tasks is increased.

Insert Table 6 about here

As was required for the success of our project, we managed to produce a wide range of variation in workload. The proportion of trial blocks with workload over redline increases from 1% in the lowest condition to 81% when all tasks at their maximum difficulty are included. With this knowledge that workload was indeed manipulated in our study, we can now turn to the major objective of the study, developing neural network models for predicting workload levels.

Neural Network Models

Our basic modeling approach was to compare a linear (one layer) perceptron to multi-layer perceptron (MLP) models that have one or more layers of processing nodes between input nodes and the output nodes. The approach to learning in this model was to use the back-propagation rule. The NN models were trained using the performance data, the condition information, and the subjective judgments of workload previously described.

We followed three steps involved in defining, training, and testing such models:

(1) An appropriate architecture was selected. The inputs of the models were determined by the task conditions and performance data. The number of hidden nodes, however, required some experimentation with alternatives. This is because too many hidden nodes leads to exact learning of the training patterns with little success at generalizing to new patterns. Too few hidden nodes can lead to failure to learn. However, it is useful to determine how models with no or few hidden nodes compare to more complex models, therefore, we examined linear models (no hidden nodes) for comparison, and we experimented with variations in the number of hidden nodes to develop some idea of the complexity of the problem space.

(2) Next, the rate of learning and the number of training epochs was selected for each model. Some common values for learning rate found in the literature which were used as a starting point, but we also varied the number of training epochs because models sometimes generalize to new cases better with less training even though the learning performance tends to improve monotonically with the number of training epochs. A subset of the data patterns were used to train the model. More specifically, data from days 3, 5, 7, and 9 were used to train the model.

(3) Finally, the models were tested for generalization by presenting new patterns of the same type as the training patterns but ones that were not used in training (test patterns came from days 4, 6, 8, and 10). Performance of the models on new patterns reflects the extent to which general properties have been learned in contrast to learning the specifics of the training patterns.

As mentioned in the previous section, our modeling efforts came to focus on models with a single output variable, i.e., the redline variable. This variable takes on two discrete values corresponding to whether workload is below redline (40 or .4 on our modified scale) or not. With this output variable, our models are essentially trying to learn to predict whether workload is at or above redline in a given situation. We will present models with different inputs, some characterizing which conditions were in effect in a particular block of trials, and some characterizing performance in the tasks being performed.

Table 7 summarizes the results with several models when the training and testing data includes data from all participants. Table 8 shows the result of developing different models for different participants. In these tables, the Model Columns show how many input, hidden(if any), and output units are in each model. All models have only one output corresponding to the redline variable. Also shown are the total number of epochs the model was trained, the number of epochs corresponding to the best Percent Correct in Training, the best Percent Correct predictions of the redline value and the best correlation between the actual and model produced outputs. For the test data, the number of epochs of training to produce the best percent correct for the test data, the best percent correct predictions of the redline value, and the best correlation between the actual and the model produced outputs for the test data. Recall that training data were taken from Days 3, 5, 7, and 9 while the test data come from Days 4, 6, 8, and 10.

Although the models vary from 66.9 to 96.0% correct predictions of the redline value for the training data, the variation is much smaller for the test data (60.2 to 75.7%). In general there seems to be little advantage for adding hidden units to the models for the percent correct in the test data. Hidden units often lead to substantial increases in performance on the training data, but only lead to 1 to 2% improvement for the test data. This result suggests that linear models do as well as more complex ones with these data. Of course, it was necessary to ensure that the coding of variables would potentially map onto a monotonic relation between the magnitude of the variable and the degree of workload. This was all done on an a priori basis, however. As the last few lines in the table show, models based on only 2 or three inputs from performance data do about as well as models with many more inputs. About the best the models can do with the test data is to predict the redline variable about 75% of the time, a value which is of interest theoretically, but it is rather far from values needed for application of these models. Perhaps the performance of models tuned to individual participant data will show more promise (see Table 8).

For some individuals, a much more impressive level of performance can be achieved, reaching 94.3% correct for one particular individual. For other individuals, performance is not nearly as impressive. Still, the average performance of individually trained and tested models exceeds that of the group trained models. Again, we find that not much is gained in performance on the test data by adding hidden units to the models. Simple linear models appear to capture the bulk of information in the data. Because the linear models are essentially equivalent to linear regression models, it is not clear that anything can be gained by using the power of nonlinear models for prediction in this case. Linear models often turn out to be particularly robust (Dawes, 1979).

Insert Tables 7 & 8 about here

Conclusions

On the positive side, our experimental tasks did produce reasonably wide variations in workload, and the models show that it is possible to predict subjective workload to a significant degree using measures of task performance. For some individuals, the level of prediction was impressive indeed, approaching 95% success in predicting when workload would exceed a critical redline value. That being said, it is clear that overall, the level of prediction achieved is rather far from what would be required to use such models in an operational setting. If a system is to be able to detect workload levels, it must be more accurate than our models for it to be of real value. Clearly, more work is required to achieve acceptable levels of accuracy.

Our work with the models showed quite clearly that separate models for different individuals will probably be required to reach high levels of accuracy in prediction. Of course, it would simplify application greatly if a general model could be developed that would predict for different individuals. It is possible, however, to develop a system that could adapt to each individual in a period of training which would produce a model that could be applied to performance following the training. For that matter, it is not all that unreasonable to periodically retrain a system to take into account changes in individuals across time. The catch, of course, is to have an appropriate architecture for such a model to begin with. Where might we look to increase the accuracy of prediction?

One promising direction is to turn to physiological measure to assist in predicting workload levels. Various physiological measures have been used in previous research on the workload of pilots in various stages of flight. Heart rate is one of the most common measures used (e.g., Kakimoto, Nakamura, Tarui, Nagasawa, & Yagura, 1988; Ruffel-Smith, 1967; Wilson & Fisher, 1991), but respiration, electroencephalography (EEG), and eye blinks have also been investigated (e.g., Harding, 1987; Sterman, Schummer, Dushenko, & Smith, 1987). It would be of value to assess the utility of several physiological measures including heart rate, respiration rate, blink rate, and EEG recorded at several sites. Models could then be developed that made use of various combinations of performance and physiological measures in an attempt to find a combination of variables that would predict workload levels with the necessary level of accuracy.

References.5 References.

Arbib, M. I. (1986). Brains, Machines, and Mathematics. New York: Springer-Verlag.

Benson, A., Schvaneveldt, R., & Waag, W. (1994). *Neural Network Models of Measurement Decision Logic.* Aircrew Training Division, Armstrong Laboratory, Williams AFB, AZ.

Borg, G. (1978). Subjective aspects of physical and mental load. *Ergonomics*, 21, 215-220.

Bortolussi, M. R., Kantowitz, B. H., & Hart, S. G. (1986). Measuring pilot workload in a motion base trainer: A comparison of four techniques. *Applied Ergonomics*, 17, 278-283.

Cohen, A., Ivry, R. I., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 16, 17-30.

Cooper, G. E., & Harper, R. P., Jr. (1969). The use of pilot rating in the evaluation of aircraft handling qualities (Report No. NASA TN-D-5153). Moffett Field, CA: Ames Research Center, National Aeronautics and Space Administration, 1969.

Corwin, W. H. (1992). In-flight and postflight assessment of pilot workload in commercial transport aircraft using the subjective workload technique. *The International Journal of Aviation Psychology*, 2, 77-93.

Curran, T. , & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1), 189-202.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34. 571-582. (Reprinted in Kaneman, D, Slovic, P. & Tversky, A. (Eds.) (1982), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press).

Frensch, P. A., Buchner, A., & Lin, J. (1994). Implicit learning of unique and ambiguous serial transitions in the presence and absence of a distractor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20(3), 567-584.

Gopher, D., & Browne, R. (1984). On the psychophysics of workload: Why bother with subjective measures?

Gopher, D., & Donchin, E. (1986). Workload–An examination of the concept. In Boff, K., Kaufman, L., & Thomas, J. (Eds.), *Handbook of Perception and Human Performance*, Vol. II (pp. 41-1 to 41-49). New York: Wiley.

Harding, R. M. (1987). Human respiratory responses during high performance flight. Neuilly-sur-Seine: AGARD/NATO, AG 312.

Haskell, B. E., & Reid, G. B. (1987) The subjective perception of workload in low-time private pilots: A preliminary study. *Aviation, Space, and Environmental Medicine*, 58, 1230-1232.

Hebb, D. O. (1949). *The Organization of Behavior.* New York: Wiley and Sons.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45, 188-196.

Kakimoto, Y., Nakamura, A., Tarui, H., Nagasawa, Y., & Yagura, S. (1988). Crew workload in JASDF C-1 transport flights: I. Change in heart rate and salivary cortisol. *Aviation, Space, and Environmental Medicine*, 59, 511-516.

Lippmann, R. (1987, April). An introduction to computing with neural nets. *IEEE ASSP Magazine*. pp. 4-22.

Moray, N., Dessouky, M. I., Kijowski, B. A., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors*, 33, 607-629.

Moray, N., Johanssen, J., Pew, R. W., Rasmussen, J., Sanders, A. F., & Wickens, C. D. (1979). *Mental workload: Its theory and measurement.* New York: Plenum Press.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19, 1-32.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In Boff, K., Kaufman, L., & Thomas, J. (Eds.), Handbook of Perception and Human Performance, Vol. II (pp. 42-1 to 42-49). New York: Wiley.

Perruchet, P., & Amorim, M-A (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4), 785-800.

Plaut, D. C., Nowlan, S. J., & Hinton, G. E. (1986). Experiments on learning by back propagation. Department of Computer Science Technical Report, Carnegie-Mellon University (CMU-CS-86-126).

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20, 585-594.

Reid, G. B., & Colle, H. A. (1988). Critical SWAT values for predicting operator workload. *Proceedings of the 32nd Annual Meeting of the Human Factors Society.*

Reid, G. B., Potter, S. S., & Bressler, J. R. (1989). Subjective workload assessment technique (SWAT): A user's guide (U). Technical Report AAMRL-TR-89-023, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH.

Reid, G. B., Shingledecker, C. A. & Eggemeier, F. T. (1981) Application of conjoint measurement to workload scale development. *Proceedings of the Human Factors Society Twenty-Fifth Annual Meeting*, 522-526.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General,* 124(2), 207-231.

Rosenblatt, F. (1962). *Principles of Neurodynamics.* New York: Spartan Books.

Ruffel-Smith, A. H. (1967). Heart rate of pilots flying aircraft on scheduled airline routes. *Aerospace Medicine*, 38, 1117-1119.

Rumelhart, D. E. & Zipser, D. (1986). Feature discovery by competitive learning. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Volume 1: Foundations. Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, Volume 1: Foundations. Cambridge: MIT Press.

Schneider, W., & Detweiler, M. (1988). The role of practice in dual-task performance: Toward workload modeling in a connectionist/control architecture. *Human Factors*, 30, 539-566.

Schvaneveldt, R. W. (1969). Effects of complexity in simultaneous reaction-time tasks. *Journal of Experimental Psychology, 81*, 289-296.

Schvaneveldt, R. W., & Chase, W. G. (1969). Sequential effects in choice reaction time. *Journal of Experimental Psychology, 80*, 1-8.

Schvaneveldt, R., Benson, A., Goldsmith, T., & Waag, W. (1992). *Neural Network Models of Air Combat Maneuvering*. AL-TR-1991 (DF10), Aircrew Training Division, Armstrong Laboratory, Williams AFB, AZ.

Schvaneveldt, R. W., & Gomez, R. L. (1996). Attention and modes of learning: Evidence from probabilistic sequences. Paper presented at the 37th annual meetings of the Psychonomic Society, Chicago, November.

Schvaneveldt, R. W., & Gomez, R. L. (1998). Attention probabilistic sequence learning. *Psychological Research*, xx, xx-xx.

Sterman, B., Schummer, G., Dushenko, T., & Smith, J. (1987). Electroencephalographic correlates of pilot performance: Simulation and flight studies. In K. Jessen (Ed.), Electrical and magnetic activity in the central nervous system: Research and clinical applications in aerospace medicine. Neuilly-sur-Seine: AGARD/NATO, CP 432.

Tsang, P. S., & Vidulich, M. A. (1994). The roles of immediacy and redundancy in relative subjective workload assessment. *Human Factors*, 36, 503-513.

Vidulich, M. A., Ward, G. F., & Schueren, J. (1991). Using the subjective workload dominance (SWORD) for projective workload assessment. *Human Factors,* 33, 677-691.

Wickens, C. D. (1986). The effects of control dynamics on performance. In Boff, K., Kaufman, L., & Thomas, J. (Eds.), *Handbook of Perception and Human Performance,* Vol. II (pp. 39-1 to 39-60). New York: Wiley.

Wickens, C. D., & Yeh, Y. Y. (1982). The dissociation of subjective ratings and performance. *Proceedings of the IEEE International Conference on Cybernetics and Society,* 584-587.

Wickens, C. D., Hyman, F., Dellinger, J., Taylor H., & Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics*, 29, 1371-1383.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35, 263-281.

Wilson, G. F., & Fisher, F. (1991). The use of cardiac and eye blink measures to determine flight segment in F4 crews. *Aviation, Space, and Environmental Medicine*, 62, 959-962.

Wolfe, J. D. (1978). Crew workload assessment: Development of a measure of operator workload (Report No. AFDL-TR-78-165). Wright-Patterson Air Force Base, Ohio: Air Force Flight Dynamics Laboratory, December 1978.

Table 1

Pairwise Correlations (|r| > .15) for the Variables in the Study

	track	seq	tone	trker	trkvr	seqrt	seqer	toner	numov	anyov	tmwl	efwl	stwl	wlsm	wlg	wli	rl
TRACK	1			.76	.37				.39	.35	.40	.35	.36	.47	.47	.45	.38
SEQ		1				.85	.42		.39	.26	.34	.35	.18	.36	.32	.31	.24
TONE			1			.48	.22	.23	.33	.23	.38	.43	.26	.45	.42	.41	.32
TRKER	.76			1	.44	.27	.29		.60	.44	.54	.34	.42	.55	.55	.52	.51
TRKVR	.37			.44	1				.56	.44	.24	.26	.16	.28	.26	.33	.29
SEQRT		.85	.48	.27		1	.49		.62	.41	.45	.44	.27	.49	.44	.45	.38
SEQER		.42	.22	.29		.49	1		.60	.37	.31	.24	.28	.35	.35	.36	.32
TONER			.23					1	.22								
NUMOV	.39	.39	.33	.60	.56	.62	.60	.22	1	.70	.49	.42	.42	.56	.56	.61	.55
ANYOV	.35	.26	.23	.44	.44	.41	.37		.70	1	.40	.40	.30	.46	.44	.46	.39
TMWL	.40	.34	.38	.54	.24	.45	.31		.49	.40	1	.64	.35	.85	.74	.63	.47
EFWL	.35	.35	.43	.34	.26	.44	.24		.42	.40	.64	1	.35	.84	.72	.68	.47
STWL	.36	.18	.26	.42	.16	.27	.28		.42	.30	.35	.35	1	.70	.85	.81	.78
WLSM	.47	.36	.45	.55	.28	.49	.35		.56	.46	.85	.84	.70	1	.97	.89	.71
WLG	.47	.32	.42	.55	.26	.44	.35		.56	.44	.74	.72	.85	.97	1	.92	.79
WLI	.45	.31	.41	.52	.33	.45	.36		.61	.46	.63	.68	.81	.89	.92	1	.84
RL	.38	.24	.32	.51	.29	.38	.32		.55	.39	.47	.47	.78	.71	.79	.84	1

Table 2

Description of the Variables in Table 1

Variable	**Description**
TRACK	Presence and level of Tracking Task
SEQ	Presence and level of SRT Task
TONE	Presence of Tone Counting Task
TRKER	Tracking Error
TRKVR	Tracking Variance
SEQRT	SRT Reaction Time
SEQER	SRT Error Rate
TONER	Tone Counting Error
NUMOV	Number of Performance Measures over Mean (no)
ANYOV	Are any Performance Measures over Mean? (ao)
TMWL	Time Workload Rating
EFWL	Effort Workload Rating
STWL	Stress Workload Rating
WLSM	Sum of tmwl, efwl, stwl

WLG	Group Scaled Workload
WLI	Individual Scaled Workload
RL	Redline: Does Workload exceed 40?

Table 3

Tracking Error as a Function of Tracking, SRT, and Tone-Counting Conditions

	No SRT Task		**Structured SRT Task**		**Random SRT Task**		
	No Tones	**Tones**	**No Tones**	**Tones**	**No Tones**	**Tones**	**Means**
No Tracking	x	x	x	x	x	x	x
Slow Tracking	0.05	x	0.09	0.10	0.09	0.10	0.09
Medium Tracking	0.09	x	0.14	0.15	0.14	0.15	0.13
Fast Tracking	0.16	x	0.21	0.22	0.22	0.24	0.21
Means	0.10	x	0.15	0.16	0.15	0.16	

Table 4

Reaction Time in the SRT Task as a function of Tracking,

SRT, and Tone-Counting Conditions

	No SRT Task		**Structured SRT Task**		**Random SRT Task**		
	No Tones	**Tones**	**No Tones**	**Tones**	**No Tones**	**Tones**	**Means**
No Tracking	x	x	0.27	0.29	0.28	0.30	0.29
Slow Tracking	x	x	0.35	0.37	0.37	0.38	0.37
Medium Tracking	x	x	0.37	0.38	0.38	0.39	0.38
Fast Tracking	x	x	0.39	0.39	0.39	0.40	0.39
Means	x	x	0.35	0.36	0.36	0.37	

Note. Reaction times are standardized so that 1200 msec = 1.0

Table 5

Group SWAT Scale Values as a function of Tracking,

SRT, and Tone-Counting Conditions

	No SRT Task	**Structured SRT Task**	**Random SRT Task**

	No		No		No		
	Tones	**Tones**	**Tones**	**Tones**	**Tones**	**Tones**	**Means**
No Tracking	x	x	0.12	0.24	0.11	0.27	0.19
Slow Tracking	0.06	x	0.27	0.45	0.29	0.48	0.31
Medium Tracking	0.16	x	0.35	0.54	0.38	0.56	0.40
Fast Tracking	0.36	x	0.49	0.65	0.54	0.67	0.54
Means	0.19	x	0.31	0.47	0.33	0.50	

Note. The SWAT values from 0 to 100 are converted to a 0 to 1 scale.

Table 6

Proportion of Workload Levels exceeding "Redline"

as a function of Tracking, SRT, and Tone-Counting Conditions

	No SRT Task		Structured SRT Task		Random SRT Task		
	No Tones	Tones	No Tones	Tones	No Tones	Tones	Means
No Tracking	x	x	0.05	0.23	0.02	0.27	0.14
Slow Tracking	0.01	x	0.22	0.64	0.31	0.67	0.37
Medium Tracking	0.12	x	0.45	0.73	0.48	0.77	0.51
Fast Tracking	0.45	x	0.64	0.77	0.70	0.81	0.67
Means	0.19	x	0.34	0.59	0.38	0.63	

Table 7

Neural Network Models Predicting Redline for All Participants Combined

		Training				Test		
Model	Inputs	Epochs	Best Epochs	Percent Correct	Best Corr	Best Epochs	Percent Correct	Best Corr
8-1	all but no & ao	20,000	359	83.2	.689	17	74.0	.528
10-1	all	1,820	190	82.8	.696	11	74.2	.541
8-32-1	all but no & ao	20,000	20,000	90.1	.812	11,140	75.7	.552
10-40-1	all	110,706	85,341	96.0	.919	46	75.0	.545
3-1	conditions	5,025	4	80.1	.648	4	71.9	.497
3-12-1	conditions	3,020	1,890	81.8	.672	23	72.7	.508
1-1	tracking task	771	2	74.1	.498	1	66.5	.314
1-1	srt task	99	2	69.0	.382	1	60.2	.266
1-1	tone task	105	2	70.7	.436	1	67.9	.321
5-1	performance	3,375	708	80.1	.643	11	73.6	.481
5-20-1	performance	10,357	1,464	82.0	.683	1,327	75.4	.549
1-1	track error	548	110	77.1	.578	25	70.6	.427
1-1	track variance	163	1	71.2	.429	1	63.6	.255
1-1	srt rt	98	3	70.5	.475	98	71.6	.373
1-1	srt error	2,893	796	71.0	.415	95	64.5	.284
1-1	tone error	4,302	2,458	66.9	.369	163	66.2	.258
1-1	numover	1,435	2	78.1	.602	1,000	73.0	.491
1-1	anyover	106	106	74.7	.495	1	60.2	.366
2-1	trk err + srt rt	1,705	21	80.3	.632	1,000	72.2	.499
2-8-1	trk err + srt rt	14,020	14	80.5	.648	11,000	72.9	.510

3-1	trk err + srt rt + numover	723	29	79.6	.649	3	73.9	.521
3-12-1	trk err + srt rt + numover	20,000	77	81.3	.675	1,000	75.6	.540

Table 8

Neural Network Models Predicting Redline for Individual Participants Separately

			Training				Test		
S	Model	Inputs	Epochs	Best Epochs	Percent Correct	Best Corr	Best Epochs	Percent Correct	Best Corr
1	1-1	numover	500	25	80.7	.399	25	73.9	.383
2	1-1	numover	500	36	86.4	.674	4	90.1	.659
3	1-1	numover	500	2	90.1	.241	50	94.3	.384
4	1-1	numover	500	4	76.1	.624	2	73.9	.616
5	1-1	numover	500	6	78.4	.590	5	81.8	.669
6	1-1	numover	500	6	81.8	.667	5	79.6	.668
7	1-1	numover	500	7	81.8	.642	3	75.0	.502
8	1-1	numover	500	39	79.6	.595	1	70.4	.452
		Average			**81.9**	**.554**		**79.9**	**.542**
1	10-1	all	2,000	203	80.7	.456	400	72.7	.335
2	10-1	all	2,000	1,285	96.6	.874	800	92.0	.738
3	10-1	all	2,000	2	90.9	.396	1	94.3	.271
4	10-1	all	2,000	1,004	88.6	.839	400	83.0	.677
5	10-1	all	2,000	219	85.2	.773	108	85.2	.763
6	10-1	all	2,000	15	89.8	.822	400	84.1	.724
7	10-1	all	2,000	1,775	92.1	.834	25	85.2	.759
8	10-1	all	2,000	15	85.2	.719	6	72.7	.510
		Average			**88.6**	**.714**		**83.7**	**.597**
1	10-10-1	all	5,000	2,674	94.3	.914	168	73.9	.405
2	10-10-1	all	5,000	3,134	100.0	.990	5	92.1	.720
3	10-10-1	all	5,000	4,129	96.6	.765	1	94.3	.292
4	10-10-1	all	5,000	1,745	92.1	.906	1,964	78.4	.576
5	10-10-1	all	5,000	4,687	93.2	.915	2,299	86.4	.749
6	10-10-1	all	5,000	3,316	96.6	.939	89	83.0	.723
7	10-10-1	all	5,000	1,789	94.3	.912	2	89.8	.769
8	10-10-1	all	5,000	1,394	97.7	.931	1,500	73.9	.473
		Average			**95.6**	**.909**		**84.0**	**.588**

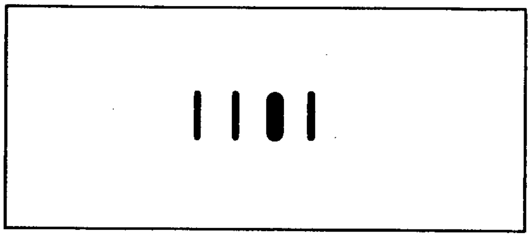


Figure 1. The stimulus display with the blot on the third position from the left.